

How to Make an Interval Version of Data Envelopment Analysis (DEA) More Adequate

Vladik Kreinovich¹, Allen J. Cutcher¹, and Minh-Tho Nguyen²

¹University of Texas at El Paso, 500 W. University
El Paso, TX 79968, USA

vladik@utep.edu, ajcutcher@miners.utep.edu

²Faculty of Finance and Banking, Ho Chi Minh City Open University
35–37 Ho Hao Hon Street, Cau Ong Lanh Ward, Ho Chi Minh City
70000, Vietnam, tho.nm@ou.edu.vn

1. What is a practical problem?

- Many companies have manufacturing facilities all around the world.
- In some places, labor is cheaper, in some places, electricity is cheaper, etc.
- Some facilities are run more efficiently, some are just starting and may need some help.
- It is therefore desirable to be able to locate facilities that need help in improving their efficiency.
- Because of the differences in costs, it is often not easy to directly compare efficiency at different locations.
- To make such a comparison possible, researchers came up with a technique known as Data Envelopment Analysis (DEA).

2. Need to take fluctuations into account

- The original DEA formulas describe each facility j by a single efficiency value E_j .
- This value provides an overall description of the facility's efficiency.
- However, it does not take into account that in real life, in general, efficiency often fluctuates.
- In some days the production is more efficient, in some days it is less efficient.
- Instead of a single efficiency value, it is desirable to come up with bounds $\underline{E}_j$ and $\overline{E}_j$.
- These bounds describe the interval $[\underline{E}_j, \overline{E}_j]$ of possible efficiency values.

3. What is known and what are the remaining problems

- Such an interval version of DEA has indeed been proposed and used.
- However, as we will show, this version does not always adequately describe the efficiency range.
- In some realistic situation it overestimates this range.
- It is therefore desirable to come up with a more economically adequate interval version of DEA.
- In this talk:
 - we explain the limitations of the current interval DEA, and
 - we describe a new, more economically adequate interval version of DEA.

4. What is DEA

- We consider the situation in which:
 - several (J) facilities produce (some or all of) M different products
 - based on N different resources: labor, electricity, components, etc.
- Let us fix a period of time – e.g., day, week, month, quarter, or year.
- To describe the efficiency of each facility, we need to introduce some notations.
- Let $y_{j,m}$ denote the amount of the m -th item produced by the j -th facility during the selected time period.
- Let $x_{j,n}$ denote the amount of the n -th resource used by the j -th facility during this time period.
- Let $u_m \geq 0$ denote the price-per-unit of the m -th item.
- Let $v_n \geq 0$ denote the cost-per-unit of the n -th resource.

5. What is DEA (cont-d)

- Under these notations, during the selected time period, the j -th facility spent the amount

$$D_j \stackrel{\text{def}}{=} \sum_{n=1}^N v_n \cdot x_{j,n}.$$

- Based on these resources, it produced the items whose overall price is

$$N_j \stackrel{\text{def}}{=} \sum_{m=1}^M u_m \cdot y_{j,m}.$$

- Thus, for each dollar spent, the facility generates the amount

$$\theta_j \stackrel{\text{def}}{=} \frac{N_j}{D_j} = \frac{\sum_{m=1}^M u_m \cdot y_{j,m}}{\sum_{n=1}^N v_n \cdot x_{j,n}}.$$

- This amount depends on the prices u_m and costs v_n .

6. What is DEA (cont-d)

- For some combinations of these prices and costs, one facility may be more efficient.
- For other combinations of prices and costs, other facilities may be more efficient.
- We say that a facility is (relatively) perfectly efficient if for some combination of the prices and costs, its productivity θ_j is the largest:

$$\theta_j = \max_k \theta_k.$$

- Sometimes, for all possible combinations of prices and costs, the facility's efficiency lacks behind others, i.e., if always

$$\theta_j < \max_k \theta_k.$$

- This means that this facility is not perfectly efficient.

7. What is DEA (cont-d)

- As a numerical value of the efficiency of the j -th facility, it makes sense to consider the largest value E_j of the ratio between:
 - its productivity and
 - the largest of the facilities' productivities:

$$E_j \stackrel{\text{def}}{=} \max_{u_1 \geq 0, \dots, u_M \geq 0, v_1 \geq 0, \dots, v_N \geq 0} \frac{\theta_j}{\max_k \theta_k}.$$

- When for some values of the prices and costs the facility is the most productive, then the value E_j is equal to 1.
- If for all possible values of the prices and costs, the j -th facility is less productive than others, then $E_j < 1$.
- This inequality is an indication to the managers that something needs to be done to increase the productivity of this facility.

8. How to compute the efficiencies

- The above expression for E_j is non-linear and sounds complicated.
- Good news is that:
 - this problem is equivalent to solving a *linear programming* (LP) problem,
 - i.e., a problem of optimizing a linear expression under linear constraints, and
 - for LP problems, there are efficient algorithms and use-friendly software implementing these algorithms.

9. How to compute the efficiencies (cont-d)

- Specifically, the problem of computing the value E_j is equivalent to solving the following linear programming problem:

Maximize $\sum_{m=1}^M u'_m \cdot y_{j,m}$ under the constraints

$$\sum_{m=1}^M u'_m \cdot y_{k,m} \leq \sum_{n=1}^N v'_n \cdot x_{k,n} \text{ for all } k$$

$$\sum_{n=1}^N v'_n \cdot x_{j,n} = 1, \text{ and}$$

$$u'_m \geq 0 \text{ and } v'_n \geq 0 \text{ for all } m \text{ and } n.$$

10. Limitations of DEA

- Many factors affect productivity.
- As a result, even when the facility remains the same, productivity changes somewhat from day to day, from year to year.
- If we simply use the production amounts during the whole period of time, we get a general information about the facility's productivity.
- However, nuances are missing.
- Some facilities may be always lacking behind, while others may be productive on some days while not very productive on others.
- These two cases require different managerial approaches.
- If a facility always lags behind, then we need to find out why, and to recommend the corresponding changes.
- On the other hand, if the facility is sometimes very efficient and sometimes not, then our problem becomes easier.

11. Limitations of DEA (cont-d)

- We just need to look at the times when the productivity was higher, and try to apply what worked then to other periods of time.
- It is therefore desirable to change the definition of the DEA so as to provide such an additional information to the managers.
- A natural idea is to show, to the managers:
 - the bounds on the efficiency, i.e.,
 - the interval $[\underline{E}_j, \overline{E}_j]$ of possible values of the efficiency of the j -th facility.

12. Current interval version of DEA

- To deal with the limitations of DEA, researchers proposed the following interval version of DEA.
- At different moments of time $t = 1, \dots, T$, the j -th facility has:
 - different amounts $x_{j,n,t}$ of resources used and
 - different amounts $y_{j,m,t}$ of items produced.
- So, the researchers suggested to use, instead of the overall values $x_{j,n}$ and $y_{j,m}$, the intervals formed by observed values:

$$[\underline{x}_{j,n}, \overline{x}_{j,n}] = \left[\min_t x_{j,n,t}, \max_t x_{j,n,t} \right] \text{ and}$$

$$[\underline{y}_{j,m}, \overline{y}_{j,m}] = \left[\min_t y_{j,m,t}, \max_t y_{j,m,t} \right].$$

13. Current interval version of DEA (cont-d)

- In this situation, as a measure of the facility's effectiveness, we take:
 - the interval $[\underline{E}_j, \overline{E}_j]$ of possible values of the ratio E_j
 - corresponding to all possible combinations of values $x_{j,n} \in [\underline{x}_{j,n}, \overline{x}_{j,n}]$ and $y_{j,m} \in [\underline{y}_{j,m}, \overline{y}_{j,m}]$.
- Interestingly:
 - the problem of computing the endpoints $\underline{E}_j$ and $\overline{E}_j$ of this interval
 - also turned out to be equivalent to an appropriate linear programming problem.
- This interval version is, thus, also feasible to solve.

14. The current interval version of DEA is not always economically adequate: an example

- We will illustrate the problem with the current interval version of DEA on the following simple example.
- Suppose that we have several identical facilities at different locations.
- We assume that during all the months – with the exception of the three summer months – each facility:
 - uses the same number of resources $x_1, \dots, x_N$ and
 - produces the same number of products $y_1, \dots, y_M$.
- In each summer month, due to vacations, production reduces by 20%.
- So, during these months, each facility uses $0.8 \cdot x_n$ resources and produces $0.8 \cdot y_m$ products.

15. From the economic viewpoint, what is the relative efficiency of each facility

- The facilities are absolutely identical and perform exactly the same way.
- Their productivities θ_j are exactly the same.
- So, for each j , for each month, we have $\theta_j = \max_k \theta_k$ and thus, $E_j = 1$.
- We therefore expect all relative efficiencies to be equal to 1, i.e., we expect to have $[\underline{E}_j, \overline{E}_j] = [1, 1]$.

16. What the current interval version of DEA returns

- Let us see what the current version of the DEA returns.
- For each j , the corresponding intervals are: $[\underline{x}_{j,n}, \bar{x}_{j,n}] = [0.8 \cdot x_n, x_n]$ and $[\underline{y}_{j,m}, \bar{y}_{j,m}] = [0.8 \cdot y_m, y_m]$.
- In this case, one can see that the largest possible values $\bar{E}_j$ of E_j is still 1.
- However, the smallest possible value $\underline{E}_j$ of the value E_j is now different from 1.
- Indeed, one can check that it corresponds to the following case:
- The efficiency θ_j of the j -th facility is the smallest possible, i.e.:
 - its used resources are the largest possible, while
 - its production is the smallest possible.

17. What the current interval version of DEA returns (cont-d)

- The efficiencies θ_k of all other facilities $k \neq j$ are the largest possible, i.e.:
 - their used resources are the smallest possible, while
 - their production amounts are the largest possible.
- So, for the j -th facility, we have $N_j = 0.8 \cdot N_j^{(0)}$.
- Here, for each quantity x , $x^{(0)}$ means its value in an idealized situation where:
 - the summer months do not differ from the rest of the year,
 - and $D_j = D_j^{(0)}$.

- Thus,

$$\theta_j = \frac{N_j}{D_j} = \frac{0.8 \cdot N_j^{(0)}}{D_j^{(0)}} = 0.8 \cdot \frac{N_j^{(0)}}{D_j^{(0)}} = 0.8 \cdot \theta^{(0)}.$$

- We skipped the index j for $\theta^{(0)}$ here, since in this situation, all the values θ are the same.

18. What the current interval version of DEA returns (cont-d)

- For all other facilities k , we have $N_k = N_k^{(0)}$ and $D_k = 0.8 \cdot D_k^{(0)}$, thus,

$$\theta_k = \frac{N_k}{D_k} = \frac{N_k^{(0)}}{0.8 \cdot D_k^{(0)}} = \frac{1}{0.8} \cdot \frac{N_k^{(0)}}{D_k^{(0)}} = 1.25 \cdot \theta^{(0)}.$$

- Thus, $\max_k \theta_k = 1.25 \cdot \theta^{(0)}$ and

$$\underline{E}_j = \frac{\theta_j}{\max_k \theta_k} = \frac{0.8 \cdot \theta^{(0)}}{1.25 \cdot \theta^{(0)}} = 0.64.$$

- This is clearly different from the economically meaningful lower end-point $\underline{E}_j = 1$.

19. Analysis of the problem

- The problem appears because of the following.
- In the current version of interval DEA, we consider all possible combinations of values $x_{j,n}$ and $y_{j,m}$.
- In reality, there may be a correlation between the values $x_{j,n}$ and $y_{j,m}$ corresponding to the same facility.
- In this case, not all pairs $(x_{j,n}, y_{j,m})$ may be possible.
- In particular, in the above example, it is not possible to have $N_k = 0.8 \cdot N_k^{(0)}$ and $D_j = D_k^{(0)}$.
- Indeed, in the months when we have $N_k = 0.8 \cdot N_k^{(0)}$, we also have $D_k = 0.8 \cdot D_k^{(0)}$.
- This is, by the way, a typical problem that happens when we naively apply interval techniques.

20. Analysis of the problem (cont-d)

- For example, for simple arithmetic operations like $y = x_1 + x_2$, $y = x_1 - x_2$, and $y = x_1 \cdot x_2$:
 - if we know the intervals $[\underline{x}_i, \bar{x}_i]$ of possible values of each input x_i ,
 - we can find the range of possible values of y :
- For addition, we have

$$[y, \bar{y}] = [\underline{x}_1 + \underline{x}_2, \bar{x}_1 + \bar{x}_2] .$$

- For subtraction, we have

$$[y, \bar{y}] = [\underline{x}_1 - \bar{x}_2, \bar{x}_1 - \underline{x}_2] .$$

- For multiplication, we have

$$[y, \bar{y}] = [\min(\underline{x}_1 \cdot \underline{x}_2, \underline{x}_1 \cdot \bar{x}_2, \bar{x}_1 \cdot \underline{x}_2, \bar{x}_1 \cdot \bar{x}_2), \max(\underline{x}_1 \cdot \underline{x}_2, \underline{x}_1 \cdot \bar{x}_2, \bar{x}_1 \cdot \underline{x}_2, \bar{x}_1 \cdot \bar{x}_2)] .$$

21. Analysis of the problem (cont-d)

- So, at first glance, to compute the range of the function $f(x) = x - x^2$ when $x \in [0, 1]$, it may seem reasonable to:
 - first use the multiplication rule to compute the range of $z \stackrel{\text{def}}{=} x^2$, and
 - then use the subtraction rule to find the desired range of $y = f(x)$.
- This procedure will lead to the following computations:

$$[\underline{z}, \bar{z}] = [0, 1] \cdot [0, 1] =$$

$$[\min(0 \cdot 0, 0 \cdot 1, 1 \cdot 0, 1 \cdot 1), \max(0 \cdot 0, 0 \cdot 1, 1 \cdot 0, 1 \cdot 1)] =$$

$$[\min(0, 0, 0, 1), \max(0, 0, 0, 1)] = [0, 1],$$

$$[\underline{y}, \bar{y}] = [0, 1] - [0, 1] = [0 - 1, 1 - 0] = [-1, 1].$$

- On the other hand, as one can easily check, the actual range of this function is much narrower: it is $[0, 0.25]$.

22. Analysis of the problem (cont-d)

- The reason why we have this problem is that the formula for the subtraction implicitly assumes that:
 - it is possible to have all possible pairs (x_1, x_2) for which each x_i belongs to the corresponding interval,
 - while in our case the values x and z are correlated: only pairs $(x, z) = (x, x^2)$ are possible.

23. Our idea

- In general, in interval techniques, the solution is to take the dependence between the variables into account.
- This is exactly what we propose.
- Instead of forming the intervals for $y_{j,m}$ and $x_{j,n}$ by combining values corresponding to different moments of time, let us:
 - use the exact values $y_{j,m,t}$ and $x_{j,n,t}$ for each moment of time t to compute the efficiencies $E_{j,t}$ at this moment of time, and
 - then combine the resulting efficiencies $E_{j,t}$ into the interval $[\underline{E}_j, \overline{E}_j]$.
- Thus, we arrive at the following method.

24. The proposed new version of interval DEA

- First, for each moment of time t and for each facility j , we compute the value

$$E_{j,t} \stackrel{\text{def}}{=} \max_{u_1 \geq 0, \dots, u_M \geq 0, v_1 \geq 0, \dots, v_N \geq 0} \frac{\theta_{j,t}}{\max_k \theta_{k,t}}.$$

- Here we denoted

$$\theta_{j,t} \stackrel{\text{def}}{=} \frac{\sum_{m=1}^M u_m \cdot y_{j,m,t}}{\sum_{n=1}^N v_n \cdot x_{j,n,t}}.$$

25. The proposed new version of interval DEA (cont-d)

- Each of these values $E_{j,t}$ can be computed by solving the following linear programming problem:

$$\begin{aligned} & \text{Maximize } \sum_{m=1}^M u'_m \cdot y_{j,m,t} \text{ under the constraints} \\ & \sum_{m=1}^M u'_m \cdot y_{k,m,t} \leq \sum_{n=1}^N v'_n \cdot x_{k,n,t} \text{ for all } k, \\ & \sum_{n=1}^N v'_n \cdot x_{j,n,t} = 1, \text{ and} \\ & u'_m \geq 0 \text{ and } v'_n \geq 0 \text{ for all } m \text{ and } n. \end{aligned} \tag{14}$$

- Once all the values $E_{j,t}$ are computed, we can compute the desired intervals $[\underline{E}_j, \overline{E}_j]$ as follows:

$$[\underline{E}_j, \overline{E}_j] = \left[\min_t E_{j,t}, \max_t E_{j,t} \right].$$

26. Comment

- For each facility j :
 - in addition to knowing the range $[\underline{E}_j, \overline{E}_j]$ of possible values of its efficiency,
 - we may want to know how frequent are different efficiency values E_j from this range.
- In other words, we may want to know the probability distribution on this interval.
- Good news is that, based on our computations, we have a sample of values $E_{j,1}, \dots, E_{j,T}$ from this distribution.
- Thus, we can use the usual statistical methods to estimate this distribution based on the sample.

27. Justification of reducing the original DEA problem to its linear programming reformulation

- Let us show that the original DEA problem is indeed equivalent to the linear programming (LP) problem.
- For this purpose, we will show that:
 - the value E_j as determined by the DEA formula
 - is equal to the maximum E'_j corresponding to the LP problem.
- Let us first prove that:
 - for each combination of values u_m and v_n ,
 - there exist values u'_m and v'_n that satisfy the LP conditions for which the sum
$$\sum u'_m \cdot y_{j,m}$$
 - that is maximized in the formula – is equal to the ratio $\theta_j / \max \theta_k$ corresponding to u_m and v_n .
- This would imply that $E'_j \geq E_j$.

28. Justification of reducing the original DEA problem to its linear programming reformulation (cont-d)

- To prove this, first, we observe that:
 - if we divide all the values u_m by the same positive constant U ,
 - i.e., replace the original values u_m with the new values $u'_m = u_m/U$,
 - then all the numerators N_k will be divide by U , i.e., their new values will be $N'_k = N_k/U$.
- Indeed, we have

$$N'_k = \sum_{m=1}^M u'_m \cdot y_{k,m} = \sum_{m=1}^M \frac{u_m}{U} \cdot y_{k,m} = \frac{1}{U} \cdot \sum_{m=1}^M u_m \cdot y_{k,m} = \frac{N_k}{U}.$$

29. Justification of reducing the original DEA problem to its linear programming reformulation (cont-d)

- Similarly:

- if we divide all the values v_n by the same positive constant V ,
- i.e., replace the original values v_n with the new values $v'_n = v_n/V$,
- then all the denominators D_k will be divide by V , i.e., their new values will be $D'_k = D_k/V$.

- Indeed, we have

$$D'_k = \sum_{n=1}^N v'_n \cdot x_{k,n} = \sum_{n=1}^N \frac{v_n}{V} \cdot x_{k,n} = \frac{1}{V} \cdot \sum_{n=1}^N v_n \cdot x_{k,n} = \frac{D_k}{V}.$$

- Let us take $U = D_j / \max \theta_k$ and $V = D_j$.
- Then, we have $D'_j = D_j/V = 1$, i.e., we have one of the LP formulas.

30. Justification of reducing the original DEA problem to its linear programming reformulation (cont-d)

- For each k , we have

$$\frac{N'_k}{D'_k} = \frac{N_k / (D_j \cdot \max_{\ell} \theta_{\ell})}{V_k / D_j} = \frac{N_k}{D_k} \cdot \frac{1}{\max_{\ell} \theta_{\ell}} = \frac{\theta_k}{\max_{\ell} \theta_{\ell}}.$$

- By definition, each value is smaller or equal than the maximum.
- Thus, we have $N'_k / D_k \leq 1$, i.e., $N'_k \leq D'_k$, which is exactly the other LP condition.
- In particular, for j , we have

$$\frac{N'_j}{D'_j} = \frac{\theta_j}{\max_k \theta_k}.$$

- Since $D'_j = 1$, we conclude that

$$N'_j = \frac{\theta_j}{\max_k \theta_k}.$$

31. Justification of reducing the original DEA problem to its linear programming reformulation (cont-d)

- So indeed:
 - the value of the above expression corresponding to the new values u'_m and v'_n
 - is equal to the ratio $\theta_j / \max \theta_k$ corresponding to the original values u_m and v_n .
- To complete our justification, let us show that, vice versa:
 - for every combination of values u'_m and v'_n that satisfy the LP conditions,
 - there exist values u_m and v_n for which the ratio $\theta_j / \max \theta_k$ is larger than or equal to the sum $\sum u'_m \cdot y_{j,m}$,
 - i.e., to the sum that is maximized in the DEA formula.
- This would imply that $E'_j \leq E_j$.

32. Justification of reducing the original DEA problem to its linear programming reformulation (cont-d)

- This, together with the inequality $E'_j \geq E_j$ proved in the first part of this justification, will imply that $E'_j = E_j$.
- Indeed, let us take $u_m = u'_m$ and $v_n = v'_n$.
- Due to LP condition, we have $N_k \leq D_k$ and thus, $\theta_k N_k / D_k \leq 1$ for all k .
- This implies that $\max_k \theta_k \leq 1$.
- Due to nother LP condition, we have $D_j = 1$, so $\theta_j = N_j / D_j = N_j$ is equal to the DEA expression.
- Since $\max \theta_k \leq 1$, we conclude that

$$\frac{\theta_j}{\sum_k \theta_k} \geq \theta_j = D_j.$$

- Thus indeed the resulting ratio $\theta_j / \max \theta_k$ is larger than or equal to the value E'_j . The justification is thus proven.

33. Acknowledgments

This work was supported by:

- the AT&T Fellowship in Information Technology,
- the Institute for Risk and Reliability, Leibniz Universitaet Hannover, Germany,
- the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Focus Program SPP 100+ 2388, Grant Nr. 501624329,
- the European Union under the project ROBOPROX (No. CZ.02.01.01/00/22 008/0004590),
- the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand,
- the Ho Chi Minh City University of Banking, Vietnam, and
- Thang Long University, Hanoi, Vietnam.